

Datenbank für Gesprochenes
Deutsch (DGD)

Schmidt, Thomas

thomas.schmidt@ids-mannheim.de
IDS Mannheim, Deutschland

Eine Korpusplattform für die Arbeit mit mündlichen Daten

Die Datenbank für Gesprochenes Deutsch (DGD) (vgl. Institut für Deutsche Sprache; Schmidt 2014a) ist die zentrale Plattform für den Zugriff auf Daten des Archivs für Gesprochenes Deutsch (AGD). Über die DGD werden 23 mündliche Korpora des Deutschen im Gesamtumfang von mehr als 3000 Stunden Audio und 8 Millionen transkribierter Wort-Tokens angeboten.

Der Bestand umfasst erstens mehrere große Variationskorpora des Deutschen, insbesondere das Korpus „Deutsche Mundarten“ (Zwerner-Korpus) mit mehreren Satelliten-Korpora nach dem gleichen Design („Deutsche Mundarten: Ehemalige deutsche Ostgebiete“, „Deutsche Mundarten: DDR“ und mehrere kleinere, regional begrenzte Sammlungen von Dialektaufnahmen) sowie das Korpus „Deutsche Umgangssprachen“ (Pfeffer-Korpus). Diese Dokumentation binnendeutscher Mundarten wird komplementiert durch Korpora auslandsdeutscher Varietäten, z. B. das Korpus „Australiendeutsch“ und drei Korpora zum Emigrantendeutsch in Israel.

Zweitens bietet die DGD Zugriff auf verschiedene Gesprächskorpora, u. a. das Berliner Wendekorpus, die Korpora „Grundstrukturen“ (Freiburger Korpus) und „Dialogstrukturen“, sowie das Korpus „Elizitierte Konfliktgespräche“. Mit dem Forschungs- und Lehrkorpus Gesprochenes Deutsch (FOLK, Schmidt 2014b) wird im AGD ein großes, breit stratifiziertes Gesprächskorpus des Deutschen aufgebaut, das technisch und methodisch auf dem Stand aktueller bester Praktiken ist und der wissenschaftlichen Gemeinschaft ebenfalls über die DGD zur Verfügung gestellt wird.

Die Bestände sind nach einem einheitlichen XML-basierten Metadatenschema dokumentiert und durch Transkriptions- und Annotationsdaten, die ebenfalls auf einem gemeinsamen XML-Datenmodell basieren, für die wissenschaftliche Analyse erschlossen.

Die DGD erlaubt zum einen ein exploratives Browsen auf diesen Daten. Korpus-, Sprecher- und Ereignisdokumentationen können eingesehen und die zugehörigen Audiodateien online abgespielt werden. Mit dem Audio alignierte Transkripte werden dem Benutzer in einer HTML5-basierten Darstellung präsentiert, die das Anspringen beliebiger Stellen im Transkript und das synchronisierte Abspielen der entsprechenden Segmente der Aufnahme ermöglicht. Diese Form des Zugangs dient sowohl dem Kennenlernen der Datenbestände als auch dem Einstieg in deren qualitative Analyse.

